Claims

[c1]   1. A computer implemented method for characterizing a plurality of biological sequences comprising:

obtaining a plurality of models, wherein each of the models represents a classification of biological sequences with structural or functional similarity;

determining fitness of the biological sequences to the models; and

automatically classifying the sequences according to the distances to the models.

[c2]   2.The method of Claim 1 wherein the plurality of biological sequences have at least 50 sequences.

[c3]   3.The method of Claim 2 wherein the plurality of biological sequences have at least 100 sequences.

[c4]   4.The method of Claim 3 wherein the plurality of biological sequences have at least 100 sequences.

[c5]   5.The method of Claim 3 wherein the models are Hidden markov models.

[c6]   6.The method of Claim 5 wherein the classification is a family and each model represents a family.

[c7]   7.The method of Claim 6 wherein the sequences are protein sequences.

[c8]   8.The method of Claim 7 wherein the distances are E-values.

[c9]   9.The method of Claim 8 wherein the step of automatically determining comprises determining a step of determining a threshold for each of the models.

[c10]   10.The method of Claim 9 wherein the step of determining a threshold comprises performing a curve analysis.

[c11]   11.The method of Claim 10 wherein the step of performing a curve analysis comprises determining a point where the e-value curve drops abruptly or flattens.

[c12]   12.A computer implemented method for gene characterization comprising:

generating libraries of models using structural relationships of known proteins;

inputting a plurality of protein sequences;

comparing the plurality of protein sequences with the models;

automatically establishing criteria for assigning the sequences for each model; and

assigning the sequences to the models based upon the criteria.

[c13]   13.The method of Claim 12 wherein the models are hidden markov models.

[c14]   14.The method of Claim 12 wherein at least 50 protein sequences are predicted protein sequences.

[c15]   15.The method of Claim 14 wherein at least 150 protein sequences are predicted protein sequences.

[c16]   16.The method of Claim 15 wherein at least 500 protein sequences are predicted protein sequences.

[c17]   17.The method of Claim 12 wherein the step of automatically establishing comprises determining a threshold for each of the models.

[c18]   18.The method of Claim 17 wherein the step of determining a threshold comprises performing a curve analysis.

[c19]   19.The method of Claim 18 wherein the step of performing a curve analysis comprises determining a point where the e-value curves drops abruptly or flattens.

[c20]   20.A system for gene annotation comprising a processor; and a memory coupled with the processor, the memory storing a plurality of machine instructions that cause the processor to perform logical steps comprising obtaining a plurality of models, wherein each of the models represents a classification of biological sequences with structural or functional similarity; determining fitness of the biological sequences to the models; and automatically classifying the sequences according to the distances to the models.

[c21] 21. The system of Claim 20 wherein the plurality of biological sequences have at least 50 sequences.

[c22] 22. The system of Claim 21 wherein the plurality of biological sequences have at least 100 sequences.

[c23] 23. The system of Claim 22 wherein the plurality of biological sequences have at least 100 sequences.

[c24] 24. The system of Claim 23 wherein the models are Hidden markov models.

[c25] 25. The system of Claim 24 wherein the classification is a family and each model represents a family.

[c26] 26. The system of Claim 25 wherein the sequences are protein sequences.

[c27] 27. The system of Claim 26 wherein the distances are E-values.

[c28] 28. The system of Claim 27 wherein the step of automatically determining comprises determining a step of determining a threshold for each of the models.

[c29] 29. The system of Claim 28 wherein the step of determining a threshold comprises performing a curve analysis.

[c30] 30. The system of Claim 29 wherein the step of performing a curve analysis comprises determining a point where the e-value curve drops abruptly or flattens.

[c31] 31. A system for gene annotation comprising a processor; and a memory coupled with the processor, the memory storing a plurality of machine instructions that cause the processor to perform logical steps comprising generating libraries of models using structural relationships of known proteins; inputting a plurality of protein sequences;
comparing the plurality of protein sequences with the models;
automatically establishing criteria for assigning the sequences for each model; and
assigning the sequences to the models based upon the criteria.

[c32]   32.The system of Claim 31 wherein the models are hidden markov models.

[c33]   33.The system of Claim 32 wherein at least 50 protein sequences are predicted protein sequences.

[c34]   34.The system of Claim 33 wherein at least 150 protein sequences are predicted protein sequences.

[c35]   35.The system of Claim 34 wherein at least 500 protein sequences are predicted protein sequences.

[c36]   36.The system of Claim 35 wherein the step of automatically establishing comprises determining a threshold for each of the models.

[c37]   37.The system of Claim 36 wherein the step of determining a threshold comprises performing a curve analysis.

[c38]   38.The system of Claim 37 wherein the step of performing a curve analysis comprises determining a point where the e-value curves drops abruptly or flattens.

[c39]   39.A computer software product of the invention comprising a computer readable medium having computer-executable instructions for performing the method comprising:

obtaining a plurality of models, wherein each of the models represents a classification of biological sequences with structural or functional similarity; determining fitness of the biological sequences to the models; and automatically classifying the sequences according to the distances to the models.

[c40]   40.The product of Claim 39 wherein the plurality of biological sequences have at least 50 sequences.

[c41]   41.The product of Claim 40 wherein the plurality of biological sequences have at least 100 sequences.

[c42]   42.The product of Claim 41 wherein the plurality of biological sequences have at least 100 sequences.

[c43] 43.The product of Claim 42 wherein the models are Hidden markov models.

[c44] 44.The product of Claim 43 wherein the classification is a family and each model represents a family.

[c45] 45.The product of Claim 44 wherein the sequences are protein sequences.

[c46] 46.The product of Claim 45 wherein the distances are E-values.

[c47] 47.The product of Claim 46 wherein the step of automatically determining comprises determining a step of determining a threshold for each of the models.

[c48] 48.The product of Claim 47 wherein the step of determining a threshold comprises performing a curve analysis.

[c49] 49.The product of Claim 48 wherein the step of performing a curve analysis comprises determining a point where the e-value curve drops abruptly or flattens.

[c50] 50.A computer software product of the invention comprising a computer readable medium having computer-executable instructions for performing the method comprising:

generating libraries of models using structural relationships of known proteins;

inputting a plurality of protein sequences;

comparing the plurality of protein sequences with the models;

automatically establishing criteria for assigning the sequences for each model; and

assigning the sequences to the models based upon the criteria.

[c51] 51.The product of Claim 50 wherein the models are hidden markov models.

[c52] 52.The product of Claim 51 wherein at least 50 protein sequences are predicted protein sequences.

[c53] 53.The product of Claim 52 wherein at least 150 protein sequences are predicted protein sequences.

[c54]   54. The product of Claim 53 wherein at least 500 protein sequences are predicted protein sequences.

[c55]   55. The product of Claim 54 wherein the step of automatically establishing comprises determining a threshold for each of the models.

[c56]   56. The product of Claim 55 wherein the step of determining a threshold comprises performing a curve analysis.

[c57]   57. The product of Claim 56 wherein the step of performing a curve analysis comprises determining a point where the e-value curves drops abruptly or flattens.